Comparable coreference expressions in parallel Czech, English, Polish and Russian data

Anna Nedoluzhko, Charles University, Prague (nedoluzko@ufal.mff.cuni.cz)
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
(maciej.ogrodniczuk@ipipan.waw.pl)

Purpose – The subject of our analysis are cohesive devices important for discourse analysis in three Slavonic languages vs. English based on translated texts, particularly coreferential expressions such as finite and infinite constructs e.g. relative clauses, participial, possessive and correlative constructions with a demonstrative pronoun. A special point of our analysis is comparison of pro-drop qualities of three Slavonic languages in comparison with English.

Design/methodology/approach – For the aim of this analysis we translated a 1000-sentence portion of the Wall Street Journal into Polish, Czech an Russian, automatically aligned them and analysed their syntactic structure using Universal Dependencies. Then we performed manual coreference annotation and analysed cohesive devices common for these languages.

Findings – The paper presents different quantitative and qualitative properties of analysed languages with respect to pro-drop characteristics. Since the texts are analysed for syntactic structure, our analysis is supported with syntactic knowledge.
Although Slavonic languages are very similar with respect to grammar and textual structure they differ significantly in many respects concerning textual coherence. The comparison with English shows e.g. that English personal pronouns in subject position usually turn into zero pronouns when moving to Slavonic languages or that the use of possessive pronouns may compensate for the absence of definite articles. At the same time the nature of the above properties of even geographically or etymologically close languages is significantly different.

Originality/Value – We begin a systematic comparison of coreference expressions among the three Slavonic languages and English which is a novel topic.

Keywords (3-5): cohesion, coreference, Slavonic languages, parallel corpus

Research type: research paper